Received: 20 June 2022 | Accepted: 6 July 2022

DOI: 10.1111/jocs.16842

REVIEW

JOURNAL OF CARDIAC SURGERY WILEY

Machine learning to predict adverse outcomes after cardiac surgery: A systematic review and meta-analysis

Jahan C. Penny-Dimri MBBS[1] | Christoph Bergmeir MSc, PhD[2] | Luke Perry MBBS[3,4] | Linley Hayes MBBS[5] | Rinaldo Bellomo MD, PhD[4,6,7,8] | Julian A. Smith MBMS, MSurgEd[1]

[1]Department of Surgery, School of Clinical Sciences at Monash Health, Monash University, Clayton, Victoria, Australia

[2]Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Monash University, Clayton, Victoria, USA

[3]Department of Anaesthesia and Pain Management, Royal Melbourne Hospital, Melbourne, Victoria, Australia

[4]Department of Critical Care, University of Melbourne, Melbourne, Victoria, Australia

[5]Department of Anaesthesia, Barwon Health, Geelong, Victoria, Australia

[6]Australian New Zealand Intensive Care Research Centre, Monash University, Melbourne, Victoria, Australia

[7]Department of Intensive Care, Royal Melbourne Hospital, Melbourne, Victoria, Australia

[8]Department of Intensive Care Research, Austin Hospital, Melbourne, Victoria, Australia

Correspondence

Jahan C. Penny-Dimri, Department of Surgery, School of Clinical Sciences at Monash Health, Monash University, Clayton, VIC, Australia.
Email: jahan.penny-dimri@monash.edu

Abstract

Background: Machine learning (ML) models are promising tools for predicting adverse postoperative outcomes in cardiac surgery, yet have not translated to routine clinical use. We conducted a systematic review and meta-analysis to assess the predictive performance of ML approaches.

Methods: We conducted an electronic search to find studies assessing ML and traditional statistical models to predict postoperative outcomes. Our primary outcome was the concordance (C-) index of discriminative performance. Using a Bayesian meta-analytic approach we pooled the C-indices with the 95% credible interval (CrI) across multiple outcomes comparing ML methods to logistic regression (LR) and clinical scoring tools. Additionally, we performed critical difference and sensitivity analysis.

Results: We identified 2792 references from the search of which 51 met inclusion criteria. Two postoperative outcomes were amenable for meta-analysis: 30-day mortality and in-hospital mortality. For 30-day mortality, the pooled C-index and 95% CrI were 0.82 (0.79–0.85), 0.80 (0.77–0.84), 0.78 (0.74–0.82) for ML models, LR, and scoring tools respectively. For in-hospital mortality, the pooled C-index was 0.81 (0.78–0.84) and 0.79 (0.73–0.84) for ML models and LR, respectively. There were no statistically significant results indicating ML superiority over LR.

Conclusion: In cardiac surgery patients, for the prediction of mortality, current ML methods do not have greater discriminative power over LR as measured by the C-index.

KEYWORDS

artificial intelligence, cardiac surgery, machine learning, meta-analysis, perioperative risk, systematic review

1 | INTRODUCTION

The high physiological demands of cardiac surgery put patients at risk of postoperative complications.[1] Cardiac surgery is, however, a data rich field, which has facilitated the development of a variety of risk stratification tools.[2] Traditionally, these data have been fitted with linear models and clinical scoring tools created to assist perioperative decision-making.[2] Novel machine learning (ML) algorithms continue to be developed with increasing complexity to fit the ever-increasing data. The nonlinearity of these algorithms may provide greater assistance in clinical decision-making and improve patient outcomes.[3]

Many ML models, such as artificial neural networks and the sequelae of deep learning strategies, can model the dimensions of a clinical predictive problem with nonlinear complexity and thereby uncover relationships and unique latent structure within the data.[4] Regardless of model design, the essential function of the ML model is to learn important features by training on a given data set to be able to make predictions or gain insights about data that was not part of the original training set.[4] These strategies are increasingly applied to clinical domains, with research promising the ability to leverage peri, and intra-operative data to predict complications and potentially improve patient care.[3,5]

The relative performance of these algorithms compared with logistic regression (LR) models, however, remains unclear. A review is needed to characterize the performance of these algorithms and determine possible future directions for this important clinical aid. We, therefore, conducted a systematic review and meta-analysis to compare the predictive performance of ML models against established methods such as LR and clinical scoring tools.

2 | METHODS

Ethics approval for this study was not required. The study was prospectively registered with PROSPERO (CRD42020196587). The Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) statement was followed.[6]

2.1 | Types of studies

This review included all original studies investigating all ML models used in the perioperative setting for predicting postoperative adverse outcomes. In particular, we included studies that assessed the performance of predicting poor outcomes. We included all observational cross-sectional, case-control, and cohort studies that compare any ML algorithm to any reference standard or a different ML algorithm. Studies were either retrospective or prospective and, if found, randomized control trials (RCTs) were included.

2.2 | Inclusion and exclusion criteria

This study considered all English language peer-reviewed studies published at any time. This study included studies on adult patients of any gender that underwent any form of cardiac surgery. Studies were excluded if they were on a pediatric population or involved cardiac transplant.

2.3 | Search strategy

We screened citations from Ovid MEDLINE (1950 to February 7th, 2022) and the Ovid Intelligent Gateway to Biomedical & Pharmacological Information (EMBASE) from inception to February 7th, 2022. The search strategy is presented in Supporting Information: Table S1. All identified articles had reference lists hand searched to identify any other possible relevant studies.

2.4 | Study selection

Abstracts and full texts were screened by two independent reviewers (Jahan C. Penny-Dimri and Linley Hayes) with conflicts resolved by a third reviewer (Luke Perry).

2.5 | Data extraction and risk of bias

Two independent reviewers (Jahan C. Penny-Dimri and Linley Hayes) assessed risk of bias using the QUADAS-2 Risk of Bias tool along with five additional previously reported signaling items that indicate bias when comparing ML models.[7,8] These two reviewers also independently extracted qualitative and quantitative data according to a data extraction template. Any disagreement was adjudicated by a third reviewer (Luke Perry). To assess discriminative performance, the concordance (C) index was collected, which corresponds to the area under the receiver operating characteristic curve (AUROC). The AUROC from any reported LR model, clinical scoring tool, and best performing ML model were collected.

2.6 | Data syntheis

After data collection, any C-indexes had their standard error calculated (SE) as described by Hanley and MacNeil.[9] Where there were sufficient studies exploring a postoperative outcome, defined as greater than five studies available, a pooled estimate was obtained and the data was presented as a forest plot. Meta-analysis was conducted using a random effects Bayesian approach.[10] To investigate any performance difference in an outcome agnostic context, a critical difference diagram was performed across all studies that reported an AUROC for a scoring tool, LR, and ML model.[11] Critical difference diagrams are a nonparametric approach, using a Wilcoxon-Holm method, that compares the performance of predictive models across different datasets.[11]

2.7 | Publication bias and sensitivity analysis

A funnel plot was generated to assess for small-study effects and publication bias.[12]

Sensitivity analysis was conducted to investigate bias by excluding studies deemed to be of high or unclear risk of bias to determine whether the conclusions from the data synthesis were robust. Sensitivity analysis was also used to investigate the role of publication year on performance. This analysis was included to test for bias from advances in ML methods over time as the field of ML is rapidly changing.

2.8 | Assessment of the evidence

The Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach was used to assess the evidence.[13]

3 | RESULTS

The search yielded 2792 references. There were 417 duplicates that were removed before abstract screening. A further 2247 references were excluded based on the title or abstract. During full-text screening a further 79 studies were excluded for a final inclusion of 51 studies (see Figure 1).

3.1 | Study characteristics

Studies were published from 1992 to 2022. The incidence of publication shows an exponential increase with time, depicted in Figure 2. Most studies published, 21 of 51, were from North America, followed by Europe with 14 studies, 12 from the Middle East or Asia, 2 from South America, and 2 from Australia.

The most common outcome was short-term mortality with 14 studies reporting 30-day mortality and 10 reporting in-hospital mortality. Acute kidney injury and prolonged mechanical ventilation were the next most common with five studies reporting these outcomes.

All studies were either retrospective (45) or prospective cohort studies (6). Most studies, 42 of 51, used either a train-test-validation split with bootstrapping or K-fold validation to generate performance metrics. Only six studies used a prospectively collected validation cohort and one study used a completely external validation cohort.

Overall, 33 of the 51 studies (65%) were at low risk of bias. Common points of bias included different features being used to train LR and ML model algorithms or no reference standard applied at all.

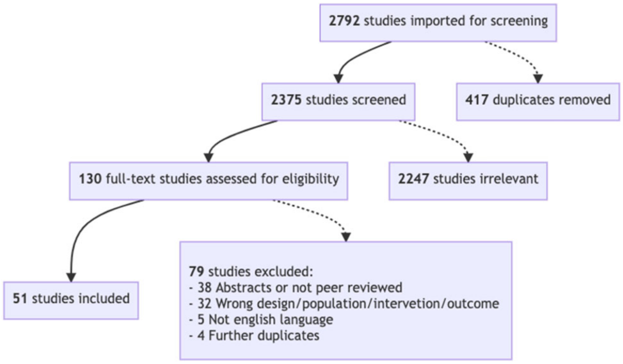
2792 studies imported for screening
2375 studies screened
417 duplicates removed
130 full-text studies assessed for eligibility
2247 studies irrelevant
51 studies included
79 studies excluded:
- 38 Abstracts or not peer reviewed
- 32 Wrong design/population/intervention/outcome
- 5 Not english language
- 4 Further duplicates

FIGURE 1 Preferred reporting items for systematic reviews and meta-analysis (PRISMA) flow diagram. The flow diagram shows the flow of studies and exclusions through the different phases of the systematic review.

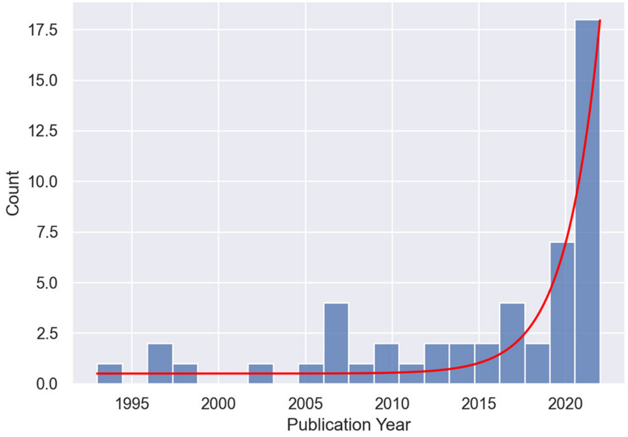

FIGURE 2 Distribution of publication rate over time for included studies. The rate of publication of applied machine learning papers in cardiac surgery is currently following an exponential distribution.

Complete tables for study characteristics are available in the Supporting Information. The descriptive study characteristics are available in Supporting Information: Table S2, the data set characteristics for each study are available in Supporting Information: Table S3, analysis characteristics for each study are available in Supporting Information: Table S4, and performance characteristics are available in Supporting Information: Table S5.

3.2 | Data synthesis

There were only sufficient study numbers to perform evidence synthesis on two outcomes, which were 30-day mortality and in-hospital mortality. Pooled C-indexes with their 95% credible intervals (CrI) are reported in Table 1. Across both 30-day mortality and in-hospital mortality, the best-performing ML model was not associated with statistically significant improvement in performance as shown by the *p*-values in Table 1. For 30-day mortality, the pooled C-index and

TABLE 1 Pooled C-indexes

Outcome	Studies	ML model pooled C-index (95% CrI)	LR pooled C-index (95% CrI)	Score pooled C-index (95% CrI)	*p*-Value ML versus LR/*p*-value ML versus score/*p*-value LR versus score
30-day mortality	14	0.82 (0.79–0.85)	0.80 (0.77–0.84)	0.78 (0.74–0.82)	0.32/0.11/0.20
In-hospital mortality	10	0.81 (0.78–0.84)	0.79 (0.73–0.84)	NA	0.24

Abbreviations: 95% CrI, 95% credible intervals; LR, logistic regression; ML, machine learning.

95% CrI were 0.82 (0.79–0.85), 0.80 (0.77–0.84), 0.78 (0.74–0.82) for ML models, LR, and scoring tool respectively. For in-hospital mortality, the pooled C-index was 0.81 (0.78–0.84) and 0.79 (0.73–0.84) for ML models and LR, respectively.

Sensitivity analysis for risk of bias and publication year did not change the results of the analysis (Supporting Information: Table S6).

Forest plots for these results are shown in Figures 3 and 4.

3.3 | Critical difference analysis

Critical difference analysis was performed on all eligible studies as well as on all low bias studies. The initial analysis showed a statistically significant difference between ML models, LR, and clinical scoring tools with ML performing the best. This was demonstrated by a *p*-value less than 0.05 across all pairwise comparisons. After removing studies at high or unclear risk of bias the difference between ML models and LR was not significant (Table 2 and Figure 5).

Additionally, sensitivity analysis for publication year showed that the initially statistically significant difference between ML and LR was lost after only including studies published after 2010, and also after restricting to studies published after 2020 (Table 2 and Supporting Information: Figures S1–S3).

3.4 | Assessment of publication bias

Visual inspection of the funnel plot showed a slight asymmetry suggesting a possibility of a small publication bias toward positive results (Supporting Information: Figure S4).

3.5 | GRADE assessment

A GRADE assessment considers five domains including the risk of bias, the precision of the estimates, the consistency, directness, and publication bias. Although our pooled estimates are precise and direct, given the high proportion of high-risk studies and small risk of publication bias, our confidence in the evidence was downgraded to low-moderate.

4 | DISCUSSION

This meta-analysis found that, in unbiased studies, the best performing ML models do not achieve a significant advantage in discriminative power compared to LR when measured by the C-index. While ML models tended toward higher performance, this was not a statistically significant result in either meta-analysis or critical difference analysis.

4.1 | Previous work

A recent Bayesian meta-analysis in 2020 investigated the use of ML for predicting mortality after cardiac surgery, however, their conclusions were that ML models outperformed LR.[3] Aside from incorporating additional evidence, an important difference between their analysis and our work is that they did not discriminate between in-hospital and 30-day mortality. Additionally, our work extends the analysis to incorporate an outcome-agnostic approach with critical difference diagrams that corroborate the findings in the main meta-analysis.

4.2 | Implications and interpretation

While modern ML methods have made significant improvements to many industries, uptake of these modeling techniques in healthcare has been slow.[14] Our findings suggest that the marginal gain in discriminative power of ML models compared to traditional statistical techniques may be part of the slow adoption of this technology. ML methods also come with trade-offs such as lack of interpretability and propensity to over-fit the training data.[15]

The current body of evidence, however, is limited with regard to assessing the true value of a predictive model. The current standard is to use metrics that assess discrimination and calibration, however, these metrics fail to provide guidance for policymakers or clinicians about the value of the model as part of a larger decision-making process.[16] Recent theoretical formulations on the ability to assess value has led to novel metrics such as area under the value operating characteristic curve, which incorporates mathematical definitions of a cost for incorrect predictions and benefit for correct predictions.[16]

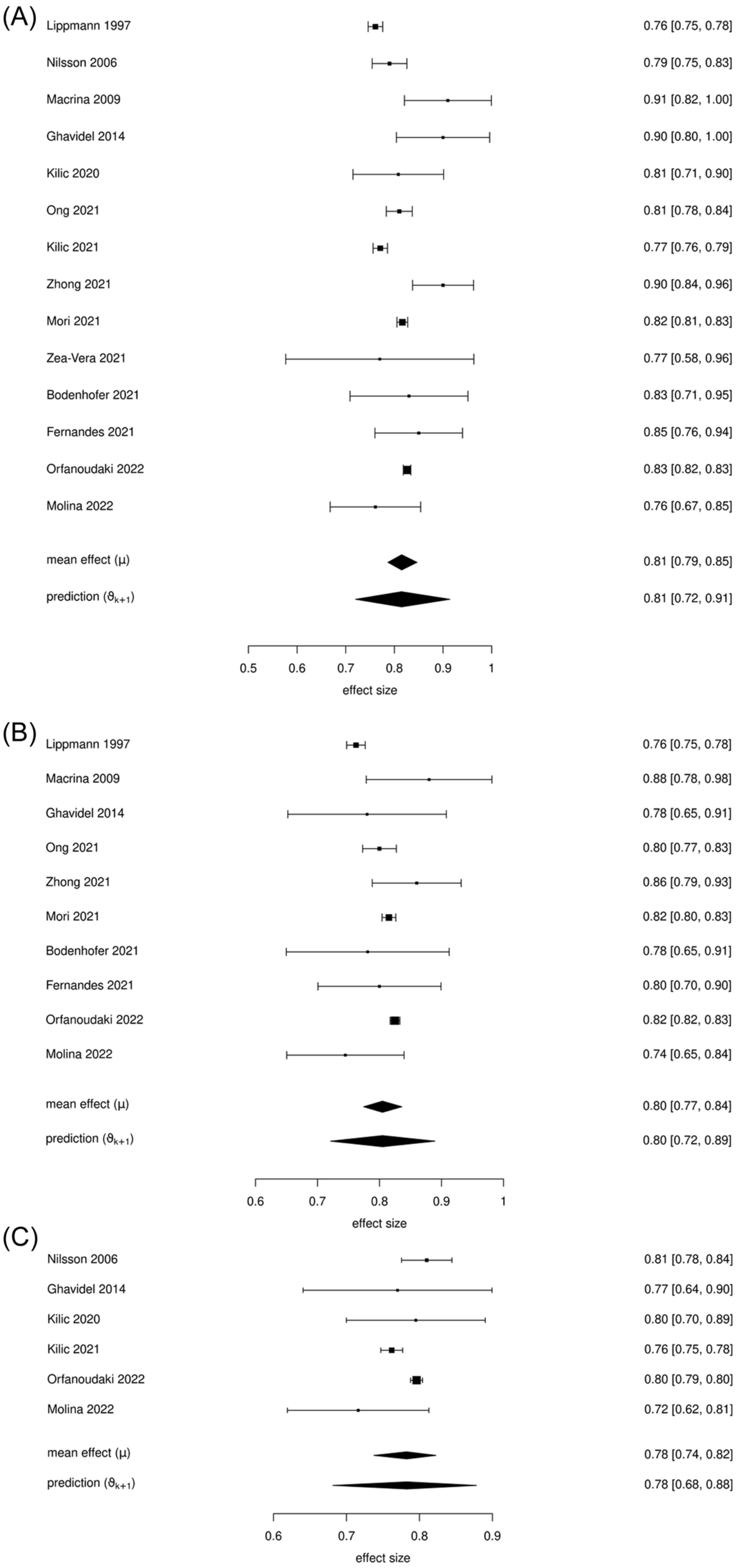

FIGURE 3 Forest plot for 30-day mortality across model type. Panel (A) shows the subgroup for the best performing machine learning model. Panel (B) shows the subgroup for logistic regression models. Panel (C) shows the subgroup for clinical scoring tools.

The studies included in this review were limited in the scope of ML modeling, incorporating mainly supervised neural networks or decision tree-based models. There are significant developments in a variety of ML methods, including unsupervised learning paradigms and unique neural network architectures that remain to be tested in this space.[5] We would therefore interpret the negative results of this study cautiously as better metrics and ML models are being rapidly developed consistent with the exponential increase in publications in this field.

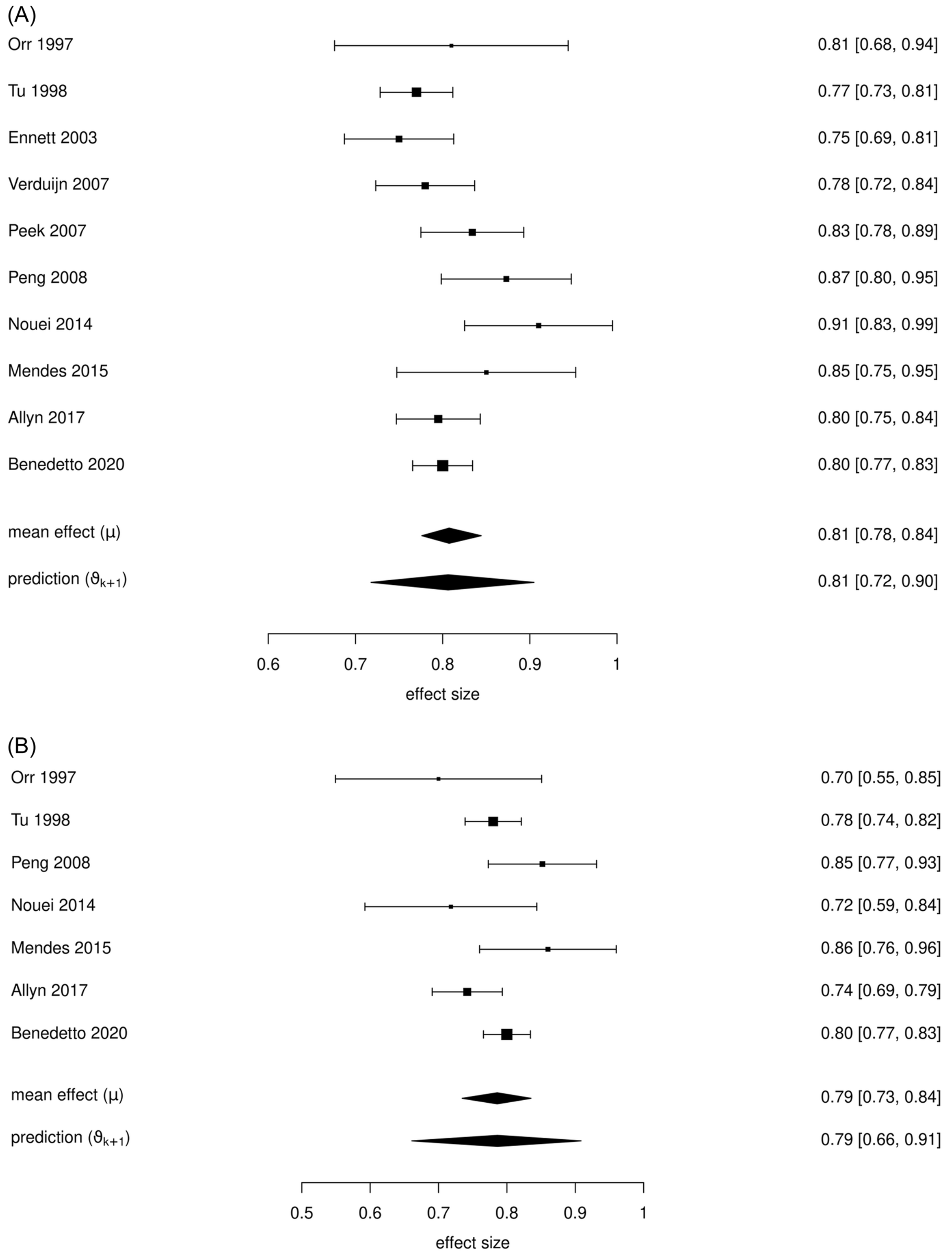

FIGURE 4 Forest plot for in-hospital mortality across model type. Panel (A) shows the subgroup for the best performing machine learning model. Panel (B) shows the subgroup for logistic regression models.

4.3 | Limitations

This study has several technical limitations. First, we were unable to perform meta-analysis on all the outcomes included in this review due to low study numbers in most of the outcomes. These low numbers also restricted the ability to perform meta-regression or subgroup analysis across import covariates, such as ML model class. Additionally, for the outcomes of 30-day mortality and in-hospital

TABLE 2 Critical difference analysis *p*-values

	ML model versus scoring tool	ML model versus LR	LR versus scoring tool
All	<0.001	0.012	<0.001
Low bias	0.002	0.052	0.004
Year > 2010	0.002	0.052	0.004
Year > 2020	0.003	0.074	0.008

Abbreviations: LR, logistic regression; ML, machine learning.

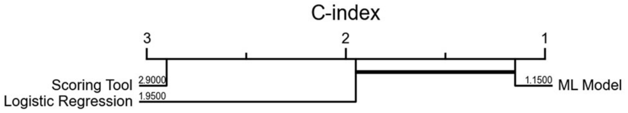

FIGURE 5 Critical difference diagram for low-risk studies. This figure compares the pairwise performance of machine learning (ML) models, logistic regression, and scoring tools and is agnostic to outcome. Groups further to the right are more performant, however, a solid line between groups indicates a lack of statistical significance. In this plot, although there is a trend toward ML models outperforming logistic regression, it is not statistically significant. All other pairwise comparisons are statistically significant.

mortality we were unable to account for patients who died outside the measurement boundary, for example after a hospital-to-hospital transfer or in-hospital and after 30-days.

Finally, this meta-analysis focused only on pooling the C-index, however, this one metric cannot incorporate many other benefits of an ML modeling approach. Recently developed explanatory modeling of previously uninterpretable ML models, such as with Shapley additive values, have made it possible to leverage the increased complexity of ML to gain unique insights from the data.[17] Several of the included studies included ML methods to provide unique prediction level risk profiles for patients.[18,19] The ability of ML methods to provide unique explanations for predictions of risk is not a measurable outcome despite being an emerging benefit of these models.

5 | CONCLUSIONS

This study found that, in cardiac surgery patients, ML models were not superior to currently used statistical methods. These findings suggest that, until better technology is developed, the clinical utility and applicability of ML technology remain a research tool only.

AUTHOR CONTRIBUTIONS

Jahan C. Penny-Dimri: Conception, data acquisition, analysis, manuscript writing. **Christoph Bergmeir**: Conception, manuscript drafting, and critical revisions. **Luke Perry**: Data acquisition, analysis, revisions. **Linley Hayes**: Data acquisition, analysis, revisions. **Rinaldo Bellomo**: manuscript drafting, and critical revisions. **Julian A. Smith**: manuscript drafting, and critical revisions.

ACKNOWLEDGMENTS

Open access publishing facilitated by Monash University, as part of the Wiley - Monash University agreement via the Council of Australian University Librarians.

ORCID

Jahan C. Penny-Dimri http://orcid.org/0000-0001-8148-1237

REFERENCES

1. Crawford TC, Magruder JT, Grimm JC, et al. Complications after cardiac operations: all are not created equal. *Ann Thorac*. 2017;103(1):32-40. doi:10.1016/j.athoracsur.2016.10.022
2. Hote M. Cardiac surgery risk scoring systems: in quest for the best. *Heart Asia*. 2018;10(1):e011017. doi:10.1136/heartasia-2018-011017
3. Benedetto U, Dimagli A, Sinha S, et al. Machine learning improves mortality risk prediction after cardiac surgery: systematic review and meta-analysis. *J Thorac Cardiovasc Surg*. 2020;163:2075-2087. doi:10.1016/j.jtcvs.2020.07.105
4. Ripley BD. *Pattern Recognition and Neural Networks*. Cambridge University Press; 1996. doi:10.1017/cbo9780511812651
5. Deo RC. Machine learning in medicine. *Circulation*. 2015;132(20):1920-1930. doi:10.1161/circulationaha.115.001593
6. Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-p) 2015 statement. *Syst Rev*. 2015;4(1):1. doi:10.1186/2046-4053-4-1
7. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Calster BV. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12-22. doi:10.1016/j.jclinepi.2019.02.004
8. Whiting PF. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529-536. doi:10.7326/0003-4819-155-8-201110180-00009
9. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36. doi:10.1148/radiology.143.1.7063747
10. Röver C. Bayesian random-effects meta-analysis using the bayesmeta r package. *J Stat Softw*. 2020;93(6):1-51. doi:10.18637/jss.v093.i06
11. Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller PA. Deep learning for time series classification: a review. *Data Min Knowl Discov*. 2019;33(4):917-963.
12. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol*. 2005;58(9):882-893. doi:10.1016/j.jclinepi.2005.01.016
13. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336(7650):924-926. doi:10.1136/bmj.39489.470347.ad
14. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J*. 2019;6(2):94-98. doi:10.7861/futurehosp.6-2-94
15. Jiang L, Wu Z, Xu X, et al. Opportunities and challenges of artificial intelligence in the medical field: current application, emerging problems, and problem-solving strategies. *J Int Med Res*. 2021;49(3):030006052110001. doi:10.1177/03000605211000157
16. Casati F, Noël PA, Yang J. On the value of ML models. *arXiv*. Preprint posted online December 13, 2021. doi:10.48550/ARXIV.2112.06775

17. Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. *Knowl Inf Syst*. 2013;41(3):647-665. doi:10.1007/s10115-013-0679-x

18. Karri R, Kawai A, Thong YJ, et al. Machine learning outperforms existing clinical scoring tools in the prediction of postoperative atrial fibrillation during intensive care unit admission after cardiac surgery. *Heart Lung Circ*. 2021;30(12):1929-1937. doi:10.1016/j.hlc.2021.05.101

19. Penny-Dimri JC, Bergmeir C, Reid CM, Williams-Spence J, Cochrane AD, Smith JA. Machine learning algorithms for predicting and risk profiling of cardiac surgery-associated acute kidney injury. *Semin Thorac Cardiovasc Surg*. 2021;33(3):735-745. doi:10.1053/j.semtcvs.2020.09.028

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Penny-Dimri JC, Bergmeir C, Perry L, Hayes L, Bellomo R, Smith JA. Machine learning to predict adverse outcomes after cardiac surgery: a systematic review and meta-analysis. *J Card Surg*. 2022;37:3838-3845. doi:10.1111/jocs.16842